April 1, 1993

A new substitution matrix for protein sequence searches based on contact frequencies in protein structures

Sanzo Miyazawa, *Gunma University*
Robert L. Jernigan, *National Institutes of Health*

Protein Engineering vol.6 no.3 pp.267–278, 1993

A new substitution matrix for protein sequence searches based on contact frequencies in protein structures

Sanzo Miyazawa and Robert L.Jernigan[1]

Gunma University, Faculty of Technology, Kiryu, Gunma 376, Japan and [1]National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

The instabilities of the native structures of mutant proteins with an amino acid exchange are estimated by using the contact energy and the number of contacts for each type of amino acid pair, which were estimated from 18 192 residue–residue contacts observed in 42 crystals of globular proteins. They were then used to evaluate a transition probability matrix of codon substitutions and a log relatedness odds matrix, which is used as a scoring matrix to measure the similarity between protein sequences. To consider amino acid substitutions in homologous proteins, base mutation rates and the effects of the genetic code are also taken into account. The average fitness of an amino acid exchange is approximated to be proportional to the structural stability of the mutant protein, which is then approximated by the average energy change of the protein native structure expected for the amino acid exchange with neglect of the energy change of the denatured state. In global and local homology searches, this scoring matrix tends to yield significantly higher alignment scores than either the unitary matrix or the genetic code matrix, and also may yield higher alignment scores for distantly related protein pairs than MDM78. One of advantages of this scoring matrix is that the equilibrium frequencies of codons and also base mutation rates can be adjusted.

Key words: contact energies/contact frequencies/homology search/sequence comparison/substitution matrix

Introduction

The relationship between protein sequence and structure has been difficult to comprehend. In a protein, the high density of interactions makes it difficult to evaluate, on a local sequence fragment, the atomic interactions that prevail in the total structure. However, we have shown that pairwise interactions between mean points in residues can serve usefully to evaluate the overall quality of protein folds. Here we focus on a smaller problem, but use the same approach to investigate the utility of such pairwise interactions to assess the substitutions of amino acids.

It is well known that the tertiary structures of proteins are well conserved in the evolutionary process of the proteins. This is because the function of a protein must be conserved during protein evolution. The function of a protein is closely related to a region of the 3-D structure of the protein. A particular tertiary structure is essential for a protein to play its function. Therefore, amino acid mutations that make native structures unstable are generally deleterious for a host organism, are therefore eliminated from and are not established within a population. This is a selection that works at the level of protein structure. As a result, the family of sequences of related functions should be informative about the ranges of viable substitutions.

The stability of protein tertiary structures is significantly affected by amino acid substitutions in the primary sequences. The effects of amino acid replacements depend on the type of the replacements. On average, the stability of tertiary structures is less affected by substitutions among amino acids with similar physico-chemical properties than by others. Dayhoff *et al.*, (1968, 1972, 1978) compiled accumulated amino acid substitution matrices from amino acid substitutions observed in closely related protein sequences and then evaluated mutation probability matrices that correspond to a transition probability matrix for amino acid substitutions in protein evolution. Based on those data, Dayhoff *et al.* (1968, 1972, 1978) pointed out that amino acid substitutions in the evolutionary process of homologous proteins often occur among similar amino acids. This observation must be interpreted by considering the effects of the genetic code, which may favor mutations between similar amino acids, the effect of unequal base mutation rates for transversion and transition, and selections at the DNA level. However, it is clear that the results analyzed by Dayhoff *et al.* (1968, 1972, 1978) reflect a selection at the level of protein.

In this paper we estimate the average degree of destabilization of a protein structure caused by an amino acid exchange, which approximately represents the average fitness of the amino acid exchange. The degree of instability of protein native structures caused by amino acid substitutions can be measured by evaluating the change of interaction energies in protein structures. In a statistical sense, each type of amino acid is found at a particular location in the three dimensional structure of proteins; non-polar residues are more often found in the non-polar environment of the protein core and polar residues on the protein surface. Residues surrounding an amino acid in protein structures are specific to the type of amino acid. We consider a typical or average protein which satisfies statistical features observed in a large set of protein structures, and evaluate the average energy increment caused by an amino acid exchange in such an average protein. On average, the degree of instability of protein structures caused by an amino acid exchange is equal to the Boltzmann factor of the average energy increment.

In our previous study (Miyazawa and Jernigan, 1985) we estimated effective inter-residue contact energies for proteins in solution from the numbers of residue–residue contacts observed in crystal structures of globular proteins by means of a quasi-chemical approximation with an approximate treatment of the effects of chain connectivity. This empirical energy function includes solvent effects, and can provide a crude estimate of the long-range component of conformational energies. By using the amino acid contact energy and the number of contacts for each type of amino acid pair, we evaluate the average energy increment of the native structure and then the degree of the instability of the native structure caused by an amino acid exchange. An energy increase results from unfavorable interactions between a substituent and surrounding residues whose distribution corresponds to that of the original residues. Those estimates of instabilities caused by exchanging each of the 20 kinds of amino acids for any other can be regarded as a selection for each type of amino acid exchange. Thus, a transition probability matrix for codon substitutions is derived from those estimates of instabilities caused by amino acid exchanges with a simple assumption about the rule of base mutation. This transition matrix reflects the base mutation rates, and includes the effects of the genetic code and conservation against amino acid exchanges.

Because the transition probability matrix for amino acid substitutions describes the likelihood of amino acid substitutions, it can be used to measure similarities between amino acid sequences. Schwartz and Dayhoff (1978) showed by using their mutation probability matrix that the log relatedness odds matrix of 250 PAM, which is calculated from the transition matrix of 250 PAM (accepted mutations per 100 residues) and represents the preference of amino acid matches and mismatches relative to random matches and mismatches, is useful in detecting distant relationships between amino acid sequences. There are several scoring matrices that have been devised for sequence comparison. The simplest one is the unitary matrix that scores only identical amino acids. Another one, which is called the genetic code matrix, uses the minimum number of base changes needed for an amino acid substitution as the basis of weighting. Feng *et al.* (1985) devised a simple matrix called a structure−genetic matrix, taking account of the structural similarities of amino acids and the genetic code. A method that we present in this paper is similar to the method of Feng *et al.* (1985) in the sense that the physico-chemical similarities of amino acids and the genetic code are taken into account, but the method used for constructing a scoring matrix is similar to the method of Dayhoff *et al.* (1978). However, the similarities among amino acids are more systematically evaluated in the present model than in Feng *et al.* (1985). The scoring matrix calculated by the present method is compared with these other scoring matrices described above in the detection of distantly related protein sequences.

Materials and methods

Mutation process

We assume that the substitution process of codons in protein coding regions of DNA can be approximated as a temporally homogeneous Markov chain. In other words, the transition probability matrix of codon substitutions, **S**, at time t is assumed to be represented by

$$\mathrm{d}\mathbf{S}/\mathrm{d}t = \mathbf{RS} \tag{1}$$

with a solution

$$\mathbf{S}(t) = \exp\,(\mathbf{R}t) \tag{2}$$

where

$$\sum_{\beta} R_{\alpha\beta} = 0 \tag{3}$$

R is a substitution rate matrix of codons; $R_{\alpha\beta}$ is the substitution rate of an α type of codon by β. Substitutions between termination codons and amino acid codons are not considered in the present analysis. It may be reasonable to assume that the detailed balance theorem is satisfied at equilibrium in such a substitution process. That is, we assume that

$$f_\alpha R_{\alpha\beta} = f_\beta R_{\beta\alpha} \tag{4}$$

where f_α is the equilibrium frequency of the α type of codons. Thus, $\mathbf{S}(t)$ can be represented by

$$\mathbf{S}(t) = \sum_{\mu} (\, v^{t}_{\alpha\mu} \exp\,(r_\mu\, t\,)\, v_{\mu\beta}\,)\, (f_\beta \,/\, f_\alpha\,)^{1/2} \tag{5}$$

where r_μ is an eigenvalue and v_μ is a left eigenvector for a real symmetric matrix, $R_{\alpha\beta}\,(\,f_\alpha \,/\, f_\beta\,)^{1/2}$, that is,

$$\sum_{\alpha} v_{\mu\alpha}\,\{\, R_{\alpha\beta}\,(\,f_\alpha \,/\, f_\beta\,)^{1/2}\,\} = r_\mu v_{\mu\beta} \tag{6}$$

r_μ has zero or a negative value and $v_{\mu\alpha}$ is a real unitary matrix. The substitution rate matrix **R** may be separated into two factors, one of which represents selection at the level of protein structure and another term for other mechanisms:

$$R_{\alpha\beta} = p_{\alpha\beta}\, M_{\alpha\beta} \quad \text{for } \alpha \neq \beta \tag{7a}$$

$$R_{\alpha\alpha} = -\sum_{\substack{\beta- \\ \neq\alpha}} R_{\alpha\beta} \tag{7b}$$

where $p_{\alpha\beta}$ is the fitness for a substitution of codon α by codon β. $M_{\alpha\beta}$ includes all other effects of selection and mutation at the DNA level for both transcription and translation. Here it should be noted that substitutions of the first, second and third bases in triplet codons may depend on each other, because selections at transcription and translation levels are included in $M_{\alpha\beta}$. Transforming $M_{\alpha\beta}$ into

$$M_{\alpha\beta} = m_{\alpha\beta} f_\beta \tag{8}$$

equation (4) is equivalent to

$$p_{\alpha\beta}\, m_{\alpha\beta} = p_{\beta\alpha} m_{\beta\alpha} \tag{9}$$

As a result, equation (7a) becomes

$$R_{\alpha\beta} = p_{\alpha\beta}\, m_{\alpha\beta} f_\beta \quad \text{for } \alpha \neq \beta \tag{10}$$

Evaluation of substitution rate based on protein selection; $p_{\alpha\beta}$

The fitness of a specific mutant of protein over its wild type could depend on how intact the function of the protein is. The functions of proteins are closely related to their native structures, that is, their 3-D structures. Thus, one may say that the fitness of a mutant protein is determined by the stability of that structure. Of course, there are specific amino acid substitutions that do not much affect the overall protein structure but can significantly decrease its function. Substitutions at active sites and on the surfaces that interact with other proteins are such substitutions. However, those substitutions must destabilize the protein−protein or substrate−enzyme interactions, so that the mutant proteins cannot retain their full functions.

The stabilities of tertiary structures of proteins and protein−protein interactions are significantly affected by amino acid substitutions in their primary structures. Probably the vast majority of amino acid substitutions destabilize protein native structures, and therefore are deleterious and consequently eliminated from the population. The effects of amino acid substitutions on protein structures depend on the type of amino acid substitutions. Those substitutions between amino acids whose physico-chemical properties are most similar to each other are likely to have the smallest effects on protein structures. The degree of instability of protein native structures caused by amino acid substitutions can be measured by evaluating the change of interaction energies in protein structures. However, because the stability of protein structures depends on both the native state and denatured state, we must evaluate the effects of the amino acid substitutions not only on the native structure but also on the denatured state.

Here, however, we would consider amino acid substitutions in protein evolution rather than amino acid replacements in protein engineering. In protein evolution which necessarily corresponds to a large time scale, it is considered that most mutations are deleterious and amino acid substitutions observed in protein evolution are almost neutral (Kimura, 1968) in natural selection.

Therefore, in the mutation data matrices compiled by Dayhoff *et al.* (1968, 1972, 1978) amino acid substitutions found in homologous proteins are to be regarded as both directions of replacement to be permitted. This condition corresponds to detailed balance at equilibrium. In other words, the amino acid substitution process in protein evolution is assumed to be completely at equilibrium. To treat such amino acid substitutions in protein evolution, we consider a case in which amino acids in a protein are exchanged. In this case, the amino acid composition is unchanged and the change in the denatured state can be neglected in the present approximation, and so we can discuss the stability of protein structures by considering the effects of substitutions only on the native structure.

Let us assume that $2\Delta\epsilon_{ij}$ $(= \Delta\epsilon_{ij} + \Delta\epsilon_{ji})$ represents the average energy increment of the native structure, which is measured relative to the denatured state, caused by exchanging the i and j types of residues in a protein. The extent of destabilization of the native state caused by the amino acid exchange is proportional to the Boltzmann factor of this energy increment. We approximate the average fitness of an amino acid exchange as the degree of the instability of the native state of the mutant protein. In other words, the fitness, $p_{\alpha\beta}$, for an exchange of codon α and codon β is represented as follows.

$$\begin{aligned} p_{\alpha\beta} &= \exp\ (\ -\ \Delta\epsilon_{i(\alpha)j(\beta)}\ /\ kT) \\ &= p_{\beta\alpha} \end{aligned} \qquad (11)$$

where $i(\alpha)$ means the i type of amino acid whose codon is α, k is the Boltzmann's constant, and T is absolute temperature.

Evaluation of the instability of protein native structures by an amino acid exchange

In our previous study (Miyazawa and Jernigan, 1985) we derived pair-wise hydrophobicities from the numbers of residue−residue contacts observed in crystal structures of globular proteins by means of the quasi-chemical approximation with an approximate treatment of the effects of chain connectivity. A basic assumption is that the average characteristics of residue−residue contacts formed in a large number of protein crystal structures reflect actual differences of interactions among residues, as if there were no significant contribution from the specific amino acid sequence in each protein, as well as ignoring intraresidue and short-range interactions. In employing a lattice model, each residue of a protein is assumed to occupy a site in a lattice and vacant sites are regarded as being occupied by an effective solvent molecule whose size is equal to the average size of a residue. Account is then taken of the effects of the chain connectivity only insofar as it imposes a limit to the size of the system, that is, the number of lattice sites; the system was regarded to be the mixture of unconnected residues and effective solvent molecules. For example, in the denatured state, the connectivity restricts the possible number of equivalent solvent molecules that have contact with the protein. The quasi-chemical approximation, in which contact pair formation is regarded as a chemical reaction, is applied to this system to obtain equilibrium constants based on average numbers of contacts; and contact energies are directly obtained from these equilibrium constants. Representing each residue by the center of its side chain atom position, contacts among residues were defined to be those within 6.5 Å. Continuing the use of the symbols in that previous study, e_{ij} represents the contact energy between i and j types of residues, N_i is the number of the i type of residues, and N_{ii} and $2N_{ij}$ $(= N_{ij} + N_{ji})$ are the number of contacts between two residues of the i type and that between the i and the j types of residue, respectively. N_i and N_{ij} were compiled from 42 crystal structures of globular proteins including 30 monomeric proteins. The total number of residue−residue contacts, N_{rr} $(= \Sigma\ N_{ij})$, was 18 192 and the total number of residues, N_r, was 9040.

These contact energies were proven to discriminate successfully between native-like conformations and incorrectly folded conformations. In a study of five small proteins (Covell and Jernigan, 1990), lattice points were fitted to C_α positions and these points used for generations of large numbers of diverse conformations, as reflected by the occurrences of almost all non-native contact pairs. Contact energies from Miyazawa and Jernigan (1985) were used to calculate average contact energies between all residue pairs. Native contact pairs then proved to be highly favored. Also, the native conformation was always found among the best 2% of the thousands of conformations when they were ranked by their total contact energies. Ranking by hydrophobicity alone proved to be substantially less successful, with the rank of the native form determined to be only better than the 12% level. These residue−residue contact energies can clearly play a useful role in screening conformations prior to more detailed atomic conformational calculations.

We proposed a simple, empirical method, based on effective inter-residue contact energies for proteins in water and taking account of contact energy changes in both native and denatured states, to estimate the change of unfolding Gibbs free energy caused by single amino acid replacements (S.Miyazawa and R.L.Jernigan, manuscript in preparation). In this method only a factor caused by the change of amino acid composition is taken into account to estimate the contact energy change by an amino acid replacement in the denatured state. The stability changes caused by single amino acid substitutions in the tryptophan synthase α subunit and bacteriophage T4 lysozyme, which were analyzed by Yutani *et al.* (1987) and Matsumura *et al.* (1988) respectively, are estimated by this simple, empirical method. The estimates of the unfolding Gibbs free energy changes correlate well with their observed values not only for hydrophobic amino acids but also when the aromatic and charged residues are included in the correlation analysis. In the case of tryptophan synthase α subunit, the changes of hydrophobic energy estimated by Yutani *et al.* (1987) were not large enough to explain the changes of unfolding Gibbs free energy. By contrast, our method yields the same magnitude of energy as the observed values in both the cases.

The average energy increment, $2\Delta\epsilon_{ij}$, of the native structure, which is measured relative to the denatured state, caused by exchanging i and j type residues in a protein, can be approximated in terms of the contact energies as follows.

$$\begin{aligned} \Delta\epsilon_{ij} &= \Delta\epsilon_{ji} \\ &\simeq \sum_k\ ((e_{jk} - e_{ik})\ N_{ik}\ /\ N_i\ +\ (e_{ik} - e_{jk})\ N_{jk}\ /\ N_j) \end{aligned} \qquad (12)$$

N_{ik}/N_i represents the distribution of k type of residues surrounding an i type of residue; that is, it represents the mean field of residues surrounding a specific type of amino acid in the native structure of a protein. So, the left hand side of equation (12) represents the average energy increment for an amino acid exchange between the i and j types of amino acids. In equation (12) the change in the free energy of the denatured state is simply neglected. This assumption would be reasonable if the local sequence effects in the denatured state are not large, especially because the amino acid composition itself does not change. For more details of this formulation, please refer to Miyazawa and Jernigan (1985), especially their equation (4a). Note that all the

Table I. Average energy increments ($\Delta\epsilon_{ij}$) for an amino acid exchange. These values of energy are represented in kT units. This table is calculated from equation (12), with the numbers of residue−residue contacts and amino acid contact energies that were compiled and evaluated from 42 crystal structures of globular proteins by Miyazawa and Jernigan (1985); the total number of residue−residue contacts is 18 192.25, and the total number of residues is 9040

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Gln	Asn	Glu	Asp	His	Arg	Lys	Pro	
		0.82	0.50	0.74	0.61	0.59	0.51	0.78	1.22	1.65	1.79	2.10	2.51	2.63	3.26	2.96	1.26	2.39	4.50	2.39	Cys
Cys	1.02		0.17	0.19	0.18	0.29	0.27	0.99	1.64	2.52	2.51	3.02	3.44	3.62	3.98	4.11	1.99	3.29	5.63	3.43	Met
Met	-0.29	0.97		0.09	0.07	0.19	0.22	0.99	1.64	2.57	2.53	3.01	3.58	3.69	4.13	4.24	1.91	3.30	5.66	3.49	Phe
Phe	0.39	0.35	0.61		0.05	0.14	0.25	1.04	1.48	2.40	2.40	2.88	3.44	3.64	4.02	4.12	2.00	3.25	5.47	3.36	Ile
Ile	-0.13	0.67	0.45	0.75		0.08	0.21	0.84	1.24	2.11	2.10	2.54	3.05	3.22	3.53	3.66	1.71	2.93	4.99	2.94	Leu
Leu	-0.02	0.50	0.48	0.48	0.61		0.14	0.54	0.77	1.49	1.48	1.85	2.26	2.46	2.76	2.86	1.19	2.17	3.96	2.15	Val
Val	-0.19	0.41	0.36	0.41	0.41	0.54		0.40	1.01	1.60	1.54	2.00	2.37	2.49	2.85	2.91	1.19	2.19	4.26	2.30	Trp
Trp	0.82	0.04	0.25	-0.23	0.13	-0.17	1.42		0.36	0.56	0.48	0.73	0.91	0.96	1.17	1.20	0.27	0.85	2.23	0.88	Tyr
Tyr	0.40	-0.56	0.14	-0.35	-0.27	-0.39	0.00	0.84		0.19	0.20	0.28	0.49	0.61	0.81	0.80	0.32	0.58	1.36	0.43	Ala
Ala	-0.51	-0.47	-0.47	-0.45	-0.45	-0.06	-0.51	-0.34	0.34		0.04	0.08	0.19	0.25	0.49	0.35	0.31	0.29	0.91	0.19	Gly
Gly	-0.18	-0.61	-0.52	-0.58	-0.56	-0.15	-0.15	-0.35	0.19	0.43		0.06	0.15	0.19	0.34	0.28	0.24	0.25	0.82	0.14	Thr
Thr	-0.62	-0.61	-0.75	-0.59	-0.77	-0.44	-0.70	-0.27	0.18	-0.09	0.45		0.08	0.10	0.25	0.20	0.32	0.20	0.50	0.07	Ser
Ser	-0.20	-0.75	-0.53	-0.68	-0.70	-0.44	-0.29	-0.04	0.16	0.00	0.28	0.48		0.06	0.13	0.10	0.41	0.23	0.47	0.09	Gln
Gln	-0.74	-0.97	-0.85	-0.95	-0.74	-0.68	-0.83	-0.02	-0.16	-0.14	-0.07	-0.20	0.48		0.15	0.10	0.41	0.19	0.37	0.08	Asn
Asn	-0.46	-0.97	-0.71	-0.83	-0.90	-0.69	-0.79	0.12	-0.16	-0.11	0.11	-0.02	0.20	0.38		0.08	0.67	0.52	0.55	0.25	Glu
Glu	-0.64	-0.87	-0.81	-0.86	-0.80	-0.42	-0.66	-0.11	0.03	0.13	-0.09	-0.19	0.25	0.19	0.36		0.64	0.43	0.53	0.18	Asp
Asp	-0.41	-0.90	-0.62	-0.78	-0.75	-0.39	-0.64	0.09	0.04	0.15	-0.10	-0.14	0.16	0.20	0.32	0.36		0.29	1.24	0.43	His
His	-0.29	-0.86	-0.43	-0.69	-0.50	-0.59	-0.70	0.35	-0.21	-0.19	-0.13	-0.15	0.37	0.21	0.13	0.18	0.54		0.50	0.20	Arg
Arg	-0.35	-0.83	-0.79	-0.82	-0.74	-0.54	-0.26	-0.31	-0.08	0.20	0.00	0.05	0.15	0.01	-0.02	-0.06	0.11	0.65		0.46	Lys
Lys	-0.81	-1.03	-1.03	-1.01	-1.06	-0.79	-0.92	-0.12	-0.20	-0.16	0.09	-0.14	0.30	0.38	0.25	0.16	0.14	0.04	0.49		Pro
Pro	-0.65	-0.82	-0.76	-0.78	-0.67	-0.47	-0.71	-0.20	0.17	-0.14	0.25	0.24	0.15	-0.12	-0.13	-0.15	0.14	0.14	-0.14	0.56	
	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Gln	Asn	Glu	Asp	His	Arg	Lys	Pro	

Table II. A log relatedness odds matrix (lower triangle) corresponding to 250 PAM for the present model with selection (BSPSM); 250 PAM corresponds to 384.5 base substitutions, 91.5% base difference and 79.8% amino acid difference in this model

$\Delta\epsilon_{ij}$ have positive values, because the contact energies e_{ij} were evaluated from the numbers of contacts N_{ij} by assuming that the residue contacts observed in proteins are at their optimum.

Evaluation of the rate of base substitutions; $m_{\alpha\beta}$

From equation (9) and (11), $p_{\alpha\beta}$ is symmetric, so $m_{\alpha\beta}$ must be symmetric;

$$m_{\alpha\beta} = m_{\beta\alpha} \tag{13}$$

In the following, we consider the simplest case:

$$\begin{aligned} m_{\alpha\beta} &= m \qquad \text{for single base substitutions} \\ &= 0 \qquad \text{for others} \end{aligned} \tag{14}$$

Equilibrium frequencies of amino acids and codons are characteristic to a protein. However, equilibrium frequencies of amino acids are taken here to be equal to the amino acid frequencies compiled and used by Dayhoff *et al.* (1978) to evaluate a mutation probability matrix, so that we can compare our result directly with their mutation data matrix. Codon frequencies are not known in their analysis because only protein sequences were utilized to collect amino acid substitutions. We assume here that each of the degenerate codons is equally used, even though it is often not true. We are interested only in statistical properties of amino acid substitutions, so that this assumption may not pose a problem.

The equilibrium frequencies of amino acids and the transition probability matrix of amino acid substitutions are calculated from those for codons as follows:

$$f_i = \sum_{\alpha} f_\alpha \, \delta_{\alpha i} \tag{15}$$

$$f_i \, S_{ij} = \sum_{\alpha} f_\alpha \, S_{\alpha\beta} \, \delta_{\alpha i} \delta_{\beta j} \tag{16}$$

where $\delta_{\alpha i}$ is the Kronecker delta function and is 1 if the codon α corresponds to the amino acid i, and otherwise 0.

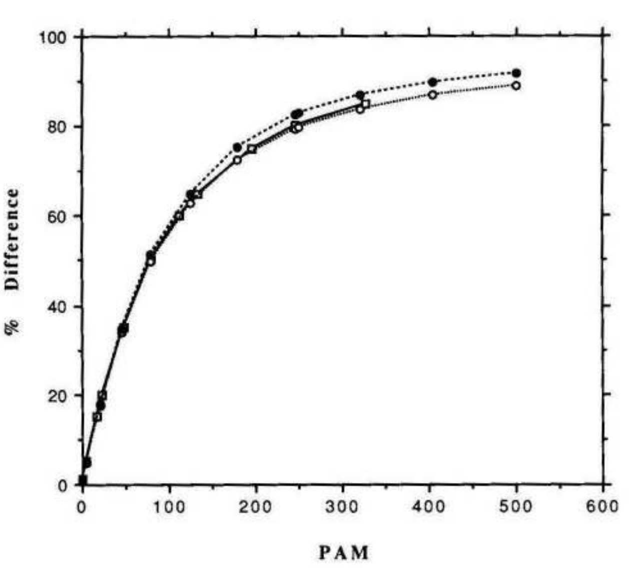

Fig. 1. The percentage difference between two sequences is plotted against the number of accepted amino acid substitutions (PAM). The solid line shows the amino acid substitution process corresponding to the mutation probability matrix compiled by Dayhoff *et al.* (1978) and the dotted line is for the present model with selection and the broken line for the present model without selection.

Results

Average energy increments of protein native structures by amino acid exchanges

Table I shows the average energy increments, $\Delta\epsilon_{ij}$, of the native structure caused by an amino acid exchange in a protein. This table displays the expected characteristics of similarities of amino acids. The most remarkable observation is the large separation between hydrophobic and hydrophilic residues. We may divide residues into three groups: a group consisting mainly of hydrophobic residues: Cys, Met, Phe, Ile, Leu, Val and Trp; a group whose member may replace both hydrophobic and

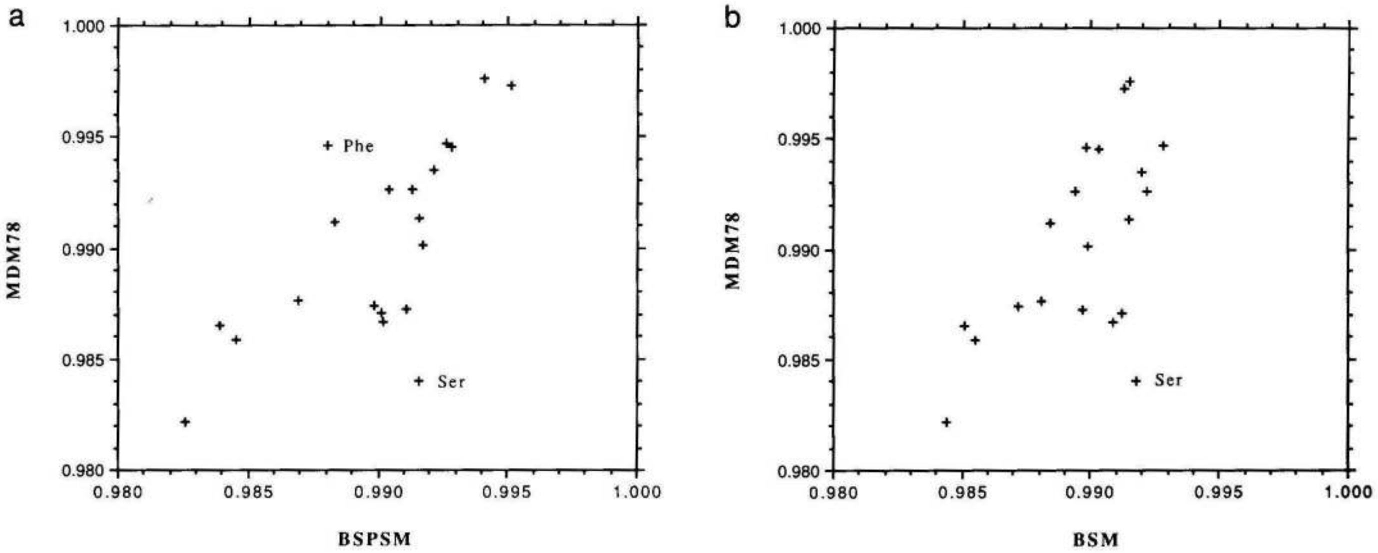

Fig. 2. Comparison of diagonal elements (S_{ii}) of the transition probability matrix corresponding to 1 PAM. The ordinate is for the mutation probability matrix compiled by Dayhoff *et al.* (1978), and the abscissa is for the present model with selection in (**a**) and without selection in (**b**).

hydrophilic residues: Tyr, Ala and His; and a group consisting mainly of hydrophilic residues: Gly, Thr, Ser, Gln, Asn, Glu, Asp, Arg, Lys and Pro. Tyr has hydrophilic characteristics as well as hydrophobic ones, probably because of the presence of a polar atom in its side chain. His can relatively easily replace Trp and Tyr, indicating the aromatic characteristics of His. Negatively charged residues, Glu and Asp, seem to be largely interchangeable. On the other hand, the characteristics of positively charged residues, Arg and Lys, are not so clear in Table I. Lys is a residue that is very unlikely to be replaced by hydrophobic residues. Replacement of Cys tends to lead to relatively large energy increments, probably because of essential Cys−Cys contacts. However, the table clearly shows the hydrophobic or buried characteristics of Cys. Also, values generally reflect the greater specificity of hydrophilic residues compared to hydrophobic ones, as we remarked earlier.

A transition probability matrix of amino acid substitutions

The transition probability matrix of codon substitutions is calculated at several time points, and the percentage difference of two amino acid sequences is calculated for each transition matrix. Figure 1 shows the substitution process of each of the mutation probability matrices (MDM78) evaluated from 1572 amino acid substitutions tabulated from closely related sequences by Dayhoff *et al.* (1978), the present model with the fitness of equation (11) for amino acid substitutions, and the present model with equal fitness for any amino acid substitution, that is, $p_{\alpha\beta} = 1$. The PAM unit is used as a time scale in Figure 1; one PAM is defined as one accepted amino acid mutation per 100 residues of protein according to the original definition. In the model with equal fitness for any amino acid substitution, base substitutions occur randomly and amino acids are substituted according to the genetic code. Clearly, MDM78 is more similar to the present model with selection than without selection, but the substitution process of MDM78 approaches equilibrium slightly more rapidly than the present model with selection. This indicates that the selection against amino acid substitutions is slightly more conservative in the present model than in MDM78. The same result is found from a comparison of the log relatedness odds matrix described in the next section.

In order to compare two models with each other in detail, the diagonal elements, S_{ii}, of the transition probability matrix of amino acid substitutions are plotted against the equivalent values of the MDM78 in Figure 2. Figure 2(a) shows the present model with selection and Figure 2(b) the case of no selection. The correlation coefficient between the first two quantities is 0.69, and this correlation is better than that in the case of no selection, where the correlation coefficient is 0.59. This relatively low correlation originates substantially from the amino acids Ser and Phe. If these two amino acids are removed from the figure, the correlation coefficient becomes 0.84 in the case of selection and 0.73 in the case of no selection; if only Ser is excluded, those correlations become 0.78 and 0.71, respectively. The mutability of Ser is significantly larger in MDM78 than in the present model, and Phe in MDM78 is less than that in the present model; the mutability of an amino acid is defined as the probability of replacing an amino acid by any other amino acid, that is, $(1 - S_{ii})$. Figure 2(b) indicates that the small mutability of Ser in the present model appears to be an effect of the genetic code.

Log relatedness odds matrix

The elements, $S(t)_{\alpha\beta}$, of the transition probability matrix of codon substitutions give the probability that codon α is changed to β at time t. On the other hand, the codon β will occur with the probability f_β in the second sequence by chance. Therefore, the log relatedness odds matrix, which was first defined and used by Dayhoff *et al.* (1978),

$$O(t)_{\alpha\beta} = \log (S(t)_{\alpha\beta} / f_\beta) \qquad (17)$$

$$= O(t)_{\beta\alpha}$$

represents the significance of the substitution between codons α and β. Codon pairs with scores above zero replace each other more often as alternatives in related sequences than in random sequences of the same amino acid composition, whereas those with scores below zero replace each other less often. This matrix may be used to evaluate the likelihood of amino acid substitutions between very distantly related sequences, and to detect evolutionary relationships between sequences. Schwartz and Dayhoff (1978) found that the log relatedness odds matrix corresponding to 250 PAM is especially appropriate for such a purpose.

Each element of the log relatedness odds matrix (BSPSM) at

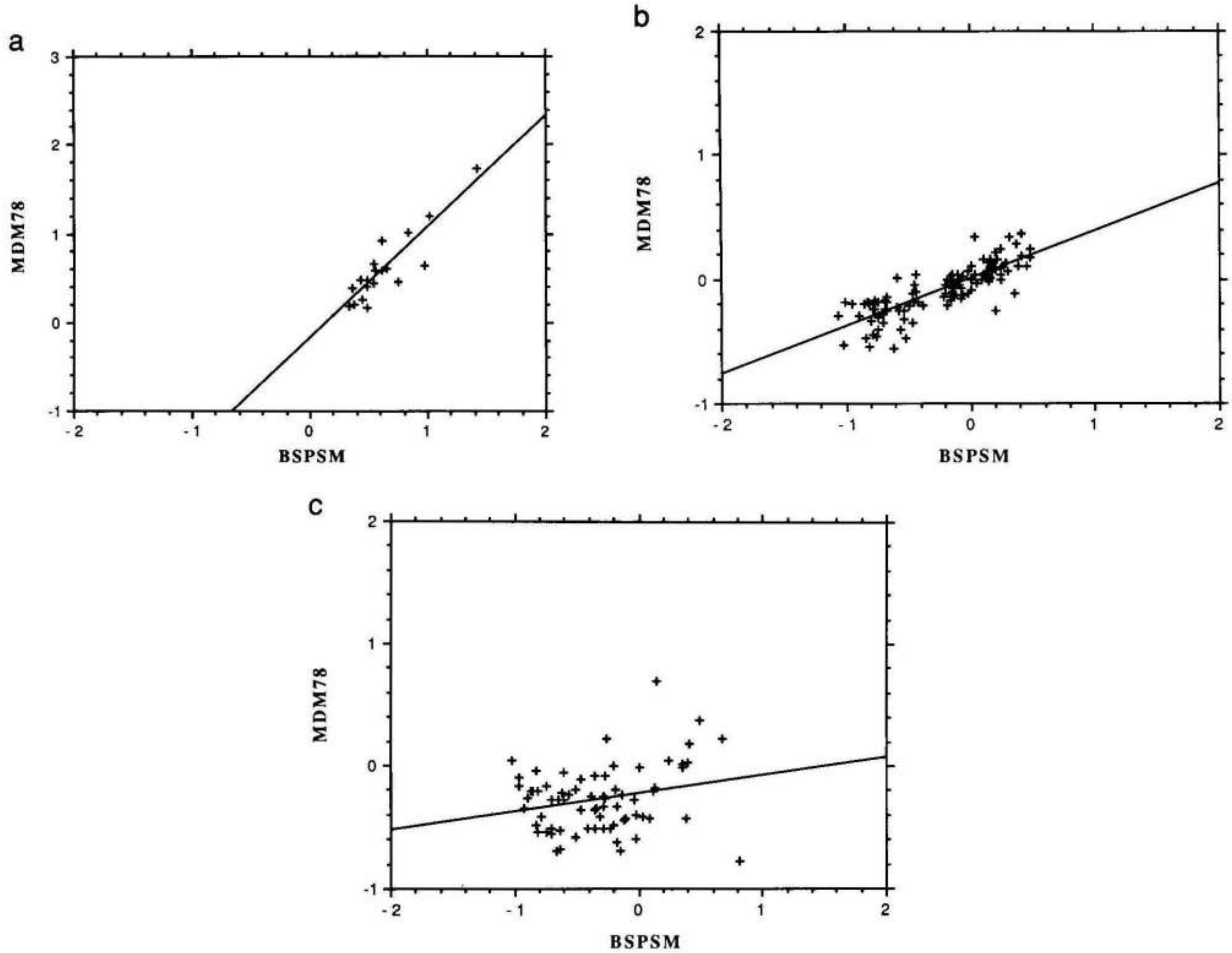

Fig. 3. Comparison of the log relatedness odds matrix corresponding to 250 PAM. The ordinate shows the values of the log relatedness odds matrix (MDM78) compiled by Dayhoff *et al.* (1978) and the abscissa shows those of the present model with selection (BSPSM). In (**a**) only diagonal elements are plotted. Off-diagonal elements are plotted in (**b**) and (**c**). In (**b**) the 70 amino acid pairs including Trp, Met, Cys and Tyr are excluded, and these others are plotted in (**c**). The solid lines in these figures are the regression lines of the ordinate on the abscissa; the regression lines are $y = -0.18 + 1.25x$ in (**a**), $y = 0.002 + 0.39x$ in (**b**) and $y = -0.23 + 0.15x$ in (**c**). The correlation coefficients are 0.89, 0.82 and 0.24 respectively.

250 PAM in the present model with selection is given in Table II, and is plotted against that of MDM78 in Figure 3. In the figure, (a) shows those diagonal elements, (b) off-diagonal elements excluding Trp, Cys, Tyr and Met, and (c) the other 70 off-diagonal elements of those amino acids. The correlation between MDM78 and the present model is good for both the diagonal elements and the off-diagonal elements excluding Trp, Met, Cys and Tyr, but it is extremely poor for all the other off-diagonal elements; the correlation coefficients are 0.89, 0.82 and 0.24 in this order for Figure 3(a,b and c). In Figure 4 we plot the values of $f_i\,(1 - S(1\mathrm{PAM})_{ii})$, which represent the proportion of each type of amino acid substitution out of the total substitutions. It is clear that the substitutions of Trp, Met, Cys and Tyr are rare, because of low frequencies and low mutabilities of those amino acids. Thus, the accumulated mutation matrix compiled from closely related sequences of Dayhoff *et al.* (1978) probably includes larger errors for these amino acids than for others. Statistically, low frequencies of those amino acids also introduce relatively large errors into the estimates of contact energies and average energy increments of amino acid exchange. These may be reasons for the poor correlation shown in Figure 3(c).

In comparison with the log relatedness odds matrices of MDM78, the matrix of the present model with selection is similar to that of MDM78 but indicates that amino acid substitutions in this model are more conservative than in the MDM78.

Use of the log relatedness odds matrix as a scoring matrix for homology search

There are several weighting schemes which have been devised for detecting similarities in comparisons between sequences. These methods all compare an amino acid in one sequence with a corresponding amino acid in another sequence, and assign a score for their similarity or difference. Consecutive segments of amino acid pairs with significantly high similarity scores are then designated as homologous alignments. The simplest scoring matrix is the unit matrix (UM), where only identical amino acids are scored. Another method uses the minimum number of bases changed as a basis for weighting. It assigns 3 for identical amino acid pairs, 2 for amino acid pairs that need at least a single base substitution to be converted to each other, 1 for amino acid pairs that need at least two base substitutions and 0 for amino acid pairs that need three base substitutions; we refer to this as the genetic code matrix (GCM). One of the most popular scoring matrices is the mutation data matrix (MDM) that was devised from accepted point mutations observed in closely related

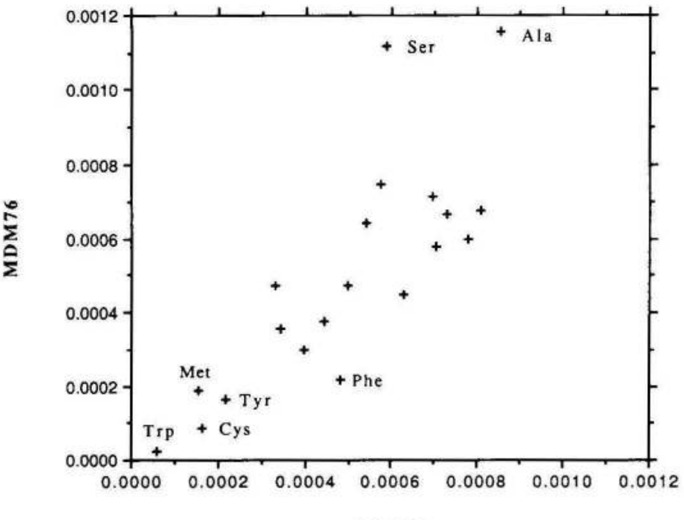

Fig. 4. Comparison of f_i $(1 - S(1\ \mathrm{PAM})_{ii})$, which is proportional to the number of substitutions of the i type of amino acids. The ordinate shows the values for the mutation probability matrix compiled by Dayhoff *et al.* (1978) and the abscissa values for the present model with selection.

Table III. Comparison of scoring matrices

(**a**) Lower triangle for correlations between off-diagonal elements of the scoring matrices and upper triangle for diagonal elements

	MDM78	AAAM	SGM	GCM	BSM	BSPSM
MDM78		0.84	–	–	0.90	0.89
AAAM	0.78		–	–	0.68	0.68
SGM	0.72	0.70		–	–	–
GCM	0.52	0.41	0.72		–	–
BSM	0.54	0.48	0.68	0.85		0.95
BSPSM	0.55	0.65	0.74	0.51	0.65	

(**b**) Correlations between off-diagonal elements of scoring matrices that do not include Trp, Met, Cys and Tyr

MDM78						
AAAM	0.84					
SGM	0.82	0.77				
GCM	0.54	0.48	0.72			
BSM	0.66	0.57	0.73	0.85		
BSPSM	0.82	0.79	0.77	0.48	0.67	
	MDM78	AAAM	SGM	GCM	BSM	BSPSM

sequences by Dayhoff *et al.* in 1967 and revised in 1969 and 1978. From a similar but different point of view, McLachlan (1971, 1972) devised a matrix based on alternative amino acids (AAAM) at each position in alignments of groups of related sequences. In a different way, Feng *et al.* (1985) devised a simple matrix called a structure–genetic matrix (SGM) that was based on the structural similarities of amino acids, as well as their likelihoods for interchanges. Since the log relatedness odds matrix of the present model is equivalent to MDM78 derived by Dayhoff *et al.* (1978), it can be used in the same way to score amino acid matches and mismatches in a sequence alignment for detecting distantly related sequences in evolution. Here, we use it as a scoring matrix for data sets of homologous sequences that are believed to have a common ancestor.

First, let us compare the log odds matrix of the present model with the other scoring matrices mentioned above. Table III shows correlation coefficients between MDM78, AAAM, SGM, GCM and the present models without selection (BSM) and with selection (BSPSM); BSM and BSPSM stand for the base-substitution matrix and base-substitution–protein-stability matrix, respectively. In Table III(a) the correlation coefficients are calculated by taking account of all amino acids pairs, but amino acid pairs including Trp, Met, Cys and Tyr are excluded in (b). The correlation of off-diagonal elements between BSM and GCM is high and the same in both cases of including or excluding infrequent amino acids. It is reasonable, because GCM may be regarded as a simple approximation of BSM and the correlation between BSM and GCM should not depend on amino acid pairs; the most probable paths of substitutions from an amino acid to another are those that require the minimum number of base substitutions, and the probability of amino acid substitutions is roughly proportional to the n power of the base mutation rate, where n is the minimum number of base substitutions needed for those substitutions, so that the log relatedness odds are proportional to the minimum number of base substitutions. The correlation between MDM78 and AAAM is reasonably good. It is expected because the probabilities of amino acid substitutions in the evolutionary process and the alternative amino acids at each position along homologous sequences are closely related to each other. Also, it is reasonable that SGM correlates well with MDM78, AAAM and BSPSM. Amino acid substitutions frequently occur between similar amino acids, and structural similarities between amino acids are taken into account in SGM. A surprising thing is that the correlations between BSPSM and both MDM78 and AAAM are as low as that between BSM and MDM78. However, the correlations between them improve markedly if amino acid pairs for infrequent amino acids with small mutabilities, Trp, Met, Cys and Tyr, are excluded in the calculation. In this case, the correlations of MDM78 with AAAM, SGM and BSPSM are similarly good. However, BSPSM correlates best with MDM78. As pointed out in the previous section, those poor correlations may result from statistical errors because of the small numbers of accumulated amino acid substitutions between those amino acid pairs in MDM78. Of course, there is also the possibility that this results from the poor estimates of amino acid interactions in protein structures because of the infrequent occurrences of those amino acids in the protein sample. However, the correlation between SGM and BSPSM does not significantly change by excluding those infrequent amino acids. This indicates that the elements of BSPSM cannot be separated into two such amino acid groups. It should be noted here that the SGM was not evaluated from actual substitution data but from the consideration of the genetic code and the similarities of amino acids. These facts indicate that the low correlation between MDM78 and BSPSM results from the statistical errors included in some elements of mutation data matrix of MDM78, or the difference in the data sets of proteins used in both the analyses.

Comparison of scoring matrices for calculating alignment scores

In this section we compare the scoring matrices, especially MDM78 and BSPSM, for global homology searches and also in local homology searches to detect distant relationships between proteins. The 'ALIGN' program, which uses a version of the Needleman and Wunsch algorithm (Needleman and Wunsch, 1970) and was written by Orcutt *et al.* (1984), is used to calculate alignment scores for global sequence alignments. The alignment score is defined to be $(s\text{-}m)/\sigma$, where s is the score for the real sequences, and m and σ are the average and the standard deviation of scores from the randomized sequences, respectively. For local homology searches, the 'LFASTA' program developed by Lipman and Pearson (1985) and Pearson and Lipman (1988) is employed to find locally homologous regions between two protein

Table IV. Alignment scores calculated with different scoring matrices using a data set originally used by Schwartz and Dayhoff (1978)

	BSPSM		MDM78		AAAM	SGM	GCM	UM
Bias	20	100	20	60	−2	0	1	0.3
Score for break	−120	−100	−80	−60	−6	−5.5	−1	−0.3
# randomized seq.	300	300	300	300	300	36	100	100
BXSMAC/ZNSMCC	2.38	0.49	3.33	1.35	2.6	2.5	3.2	3.1
FECLCP/FESG	3.83	3.78	3.82	3.66	1.8	4.0	1.6	0.1
HAHU/MYHU	7.94	7.83	10.77	10.76	9.9	15.7	6.6	5.8
HAHU/GGICE3	5.06	5.62	5.01	4.41	3.2	3.0	2.4	2.0
CCHO/CCSG6	4.11	4.07	6.36	6.01	7.3	4.7	4.3	4.5
CCHO/CCDV5M	1.74	2.30	4.93	4.24	0.4	0.2	0.4	0.2
MGHUB2/MHHU	4.14	4.00	3.56	3.40	4.7	4.0	3.3	3.6
MHHU/EHHU	7.74	7.37	12.63	10.59	9.2	8.6	9.0	4.7

The results for AAAM, GCM and UM are taken from Schwartz and Dayhoff (1978) and those of SGM are taken from Feng *et al.* (1985). The 'ALIGN' program was used to calculate alignment scores for BSPSM and MDM78. The scoring matrices of BSPSM and MDM78 are equal to the log relatedness odds matrix multiplied by 100; see Table II for BSPSM. The results for SGM may include larger statistical errors than others, because only 36 random sequences were used to estimate those alignment scores. The codes for sequences are:

BXSMAC:	antibacterial substance A; *Streptomyces* sp. (1−87)
ZNSMCC:	neocarzinostatin; *Streptomyces* sp. (1−113)
FECLCP:	ferredoxin; *Clostridium pasteurianum* (1−55)
FESG:	ferredoxin; *Spirulina maxima* (1−98)
HAHU:	hemoglobin alpha chain; human and chimpanzees (1−141)
MYHU:	myoglobin; human (1−153)
GGICE3:	globin CTT-III; midge larva (1−136)
CCHO:	cytochrome *c*; horse (1−104)
CCSG6:	cytochrome *c*6; *Spirulina maxima* (1−89)
CCDV5M:	cytochrome *c*553; *Desulfovibrio vulgaris* (1−79)
MGHUB2:	β_2-microglobulin precursor; human (21−119)
MHHU:	Ig μ chain C region; human (323−451)
EHHU:	Ig ϵ chain C region; human (320−428).

Table V. Alignment scores calculated with different scoring matrices; the global homology search of the hemoglobin superfamily originally tested by Feng *et al.* (1985)

	BSPSM		MDM78		SGM	GCM	UM
Bias	20	100	20	60	0	0	0
Score for break	−120	−100	−80	−60	−5.5	−4	−2.5
# randomized seq.	100	100	100	100	100	100	100
HBHU/HGHUA	29.36	26.32	36.55	32.88	38.2	45.4	57.5
HAHU	18.36	15.87	20.08	18.59	19.7	21.8	22.0
MYHU	6.36	7.69	8.98	9.36	9.5	7.6	8.8
GGHF3G	7.38	6.00	9.15	7.44	5.4	5.1	5.3
GGGACR	5.90	6.84	6.85	7.34	6.9	6.0	4.3
GGWNS	2.43	1.95	3.70	3.62	2.8	2.7	4.4
GGEWA3	5.66	4.21	3.78	3.99	4.0	3.8	2.8
GPFBA	6.39	6.25	6.22	6.83	7.3	4.2	1.3
HGHUA/HAHU	20.17	18.48	22.76	20.76	19.0	19.4	22.5
MYHU	8.81	8.75	9.19	9.60	12.1	8.3	8.1
GGHF3G	7.97	6.00	11.03	7.81	5.8	5.6	5.8
GGGACR	7.68	7.34	9.29	9.73	7.8	4.9	2.6
GGWNS	2.24	2.43	4.16	4.30	2.9	3.1	5.3
GGEWA3	6.57	5.00	4.19	4.35	4.4	3.4	3.1
GPFBA	4.64	5.83	5.57	6.95	6.7	4.3	3.0
HAHU/MYHU	7.63	8.22	11.88	11.60	11.5	11.0	12.9
GGHF3G	8.86	5.44	10.13	7.36	6.6	5.8	6.8
GGGACR	6.31	6.20	9.10	9.24	7.5	4.9	3.2
GGWNS	3.99	3.39	7.05	6.02	6.2	6.2	5.4
GGEWA3	4.01	4.53	6.76	7.41	5.7	3.4	0.2
GPFBA	6.52	7.95	7.02	8.04	6.4	1.5	0.4
MYHU/GGHF3G	5.45	4.15	5.29	3.99	2.9	1.9	3.3
GGGACR	6.45	6.62	6.88	6.86	7.5	8.0	3.9
GGWNS	2.59	2.82	2.56	2.19	3.6	1.8	1.2
GGEWA3	5.84	4.40	4.06	4.31	3.9	2.3	−0.5
GPFBA	3.84	4.94	4.15	5.37	4.8	2.3	1.2
GGHF3G/GGGACR	5.54	4.68	7.66	5.94	3.9	2.6	2.4
GGWNS	4.37	3.26	4.21	3.79	4.6	1.8	−0.3
GGEWA3	4.10	3.50	4.64	4.25	2.9	2.8	3.0
GPFBA	3.91	3.62	3.35	3.36	2.9	1.5	1.7
GGGACR/GGWNS	3.65	3.54	5.76	5.58	3.6	3.9	3.1
GGEWA3	3.32	3.16	4.43	4.64	5.0	3.2	2.6
GPFBA	3.30	4.89	4.42	5.18	5.7	3.5	−0.3
GGWNS/GGEWA3	13.63	10.71	15.60	15.49	13.8	13.9	15.8
GPFBA	1.51	3.07	2.89	4.14	2.9	2.2	0.2
GGEWA3/GPFBA	6.32	4.86	6.19	6.48	6.6	2.9	2.2

The 'ALIGN' program was used to calculate alignment scores for BSPSM and MDM78. The scoring matrices of BSPSM and MDM78 are equal to the log relatedness odds matrix multiplied by 100; see Table II for BSPSM. The results for SGM, GCM and UM are taken from Feng *et al.* (1985). They assigned the score 2.0 for matched cysteines in the UM and 4.0 in the GCM for amino acid substitutions that need at least three base mutations; see the reference for details. The codes of sequences are:

HBHU:	hemoglobin beta chain; human, chimpanzees and gorilla
HGHUA:	hemoglobin gamma chains; human and chimpanzee
HAHU:	hemoglobin alpha chain; human and chimpanzees
MYHU:	myoglobin; human
GGHF3G:	globin III; Atlantic hagfish
GGGACR:	globin; water snail
GGWNS:	globin, extracellular, small chain; *Tylorrhynchus heterochaetus*
GGEWA3T:	globin AIII; common earthworm
GPFBA:	leghemoglib a; kidney bean.

sequences, and the 'RDF2' program, which was also developed by them, is used to calculate alignment scores for the homologous regions. All protein sequences used in this work are taken from the PIR protein database.

Schwartz and Dayhoff (1978) tried to compare the scoring matrix of MDM78 with AAAM, GCM and UM in detecting distant relationships between proteins. Their tests were performed for global sequence alignments on a small data set of homologous proteins. Table IV lists alignment scores for those homologous proteins by using BSPSM, MDM78, AAAM, SGM, GCM and UM; the results of AAAM, GCM and UM are taken from Schwartz *et al.* (1978), and those of SGM from Feng *et al.* (1985). The 'ALIGN' program needs two additional parameters, the score for a segment of gaps and a matrix bias that is added to the scoring matrix. Two parameter sets are used for MDM78 and BSPSM: a parameter set with a large value of the matrix bias and another with a small value of the matrix bias. Small values of the matrix bias tend to yield a relatively large number of deletions and additions, so that it is appropriate for distantly related sequence pairs where there have been many changes in length.

In Table IV BSPSM fails to yield scores >3 SD for two protein pairs, but MDM78 fails only for one protein pair in one of the parameter sets and for no protein pair in the alternative parameter set. This result for BSPSM is comparable with that of SGM and better than those of AAAM, GCM and UM. An interesting result is that the alignment score of BSPSM is not so high as that of other scoring matrices for some proteins for which MDM78 and others yield large values of the alignment score. The same feature is shown in other data sets and it will be discussed later.

Feng *et al.* (1985) also tried to test score matrices by using data sets including more protein pairs than Table IV. One of them is the hemoglobin superfamily for which alignment scores are

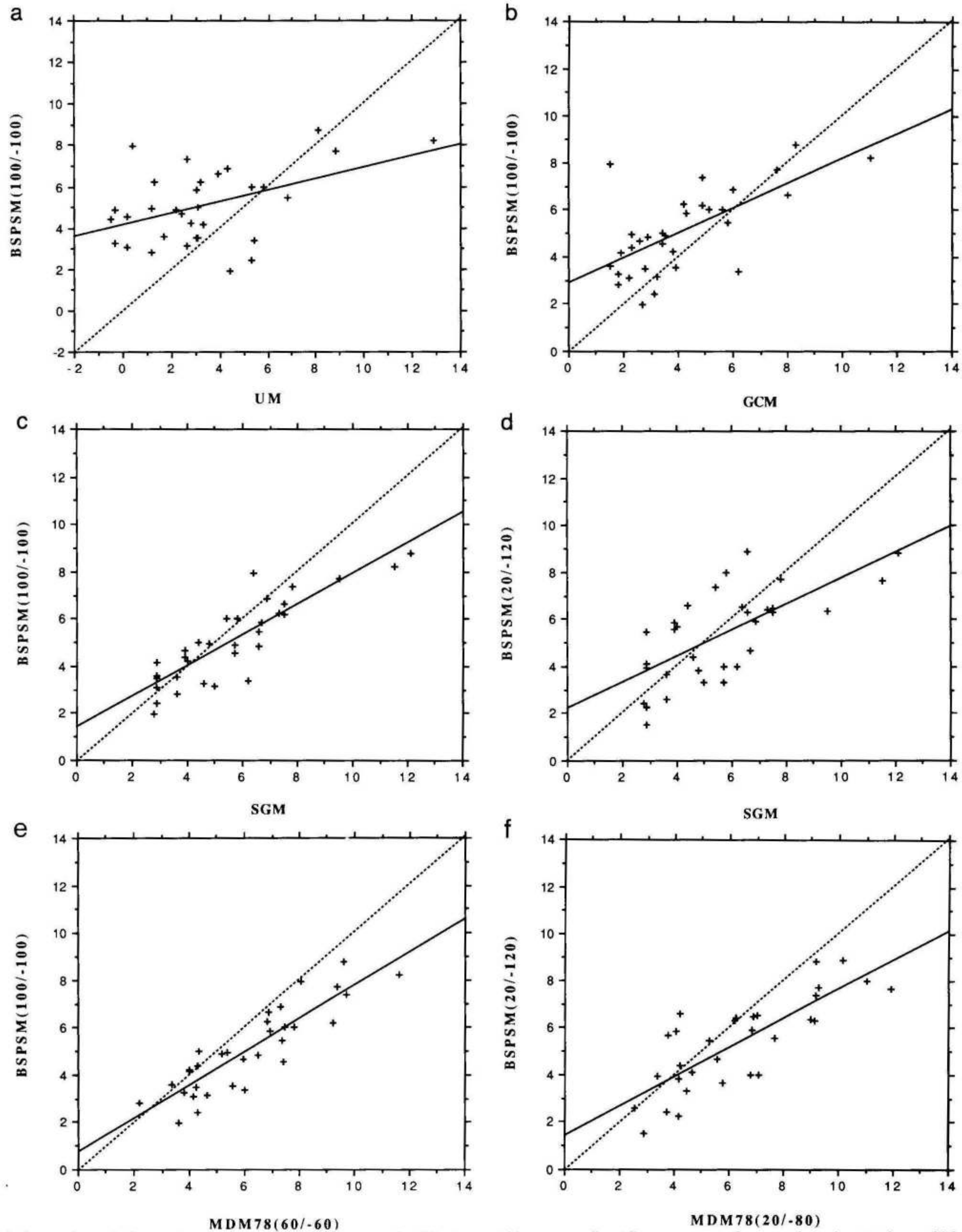

Fig. 5. Comparison of alignment scores in global homology search of the hemoglobin superfamily. Alignment scores for several scori·g matrices, which are listed in Table V, are plotted against those for other scoring matrices. The dotted lines in these figures show the isoscore line. The solid lines are the regression lines of the ordinate on the abscissa. **(a)** UM versus BSPSM with 100 for bias and −100 for break score. The regression line is $y = 4.12 + 0.28x$, and the correlation coefficient is 0.45. **(b)** GCM versus BSPSM with 100 for bias and −100 for break score. The regression line is $y = 2.85 + 0.53x$, and the correlation coefficient is 0.67. **(c)** SGM versus BSPSM with 100 for bias and −100 for break score. The regression line is $y = 1.39 + 0.65x$, and the correlation coefficient is 0.87. **(d)** SGM versus BSPSM with 20 for bias and −120 for break score. The regression line is $y = 2.18 + 0.55x$, and the correlation coefficient is 0.68. **(e)** MDM78 with 60 for bias and −60 for break score versus BSPSM with 100 for bias and −100 for break score. The regression line is $y = 0.72 + 0.70x$, and the correlation coefficient is 0.88. **(f)** MDM78 with 20 for bias and −80 for break score versus BSPSM with 20 for bias and −120 for break score. The regression line is $y = 1.41 + 0.62x$, and the correlation coefficient is 0.80.

Table VI. Comparison of scoring matrices in local homology search of the kinase-related transforming protein family

	MDM78			BSPSM		
	Identity (%)	Aligned length	Alignment score (SD)	Identity (%)	Aligned length	Alignment score (SD)
OKBO2C						
TVBE66	21.3	244	21.37	24.3	202	19.385
TVBY8	26.7	221	21.4	23	270	17.5
TVCHMS	24.3	111	13.635	23.3	146	13.795
TVHUP1	26.5	294	22.64	24.9	277	21.195
TVHURS	23.0	283	13.41	21.9	233	16.06
TVMVF6	24.9	201	11.185	23.9	138	7.34
TVBE66						
TVBY8	28.2	262	16.23	26.6	184	14.5
TVCHMS	24.1	112	14.655	29.7	101	13.575
TVHUP1	30.2	149	19.58	35.2	88	13.22
TVHURS	23.7	173	6.375	16.8	208	5.7
TVMVF6	27.8	162	17.635	28.2	156	13.5
TVBY8						
TVCHMS	32.4	102	9.685	34.4	90	12.88
TVHUP1	24.5	274	20.455	24	267	20.215
TVHURS	25.2	222	13.195	24.7	231	13.5
TVMVF6	23.0	226	13.755	24.3	185	10.1
TVCHMS						
TVHUP1	22.8	224	5.8	22.2	230	9
TVHURS	19.3	296	12.58	20.6	272	11.28
TVMVF6	24.7	215	21.15	28.2	163	19.255
TVHUP1						
TVHURS	23.7	139	4.585	21.2	165	7.22
TVMVF6	21.1	275	8.66	23.7	228	9.92
TVHURS						
TVMVF6	28.8	267	20.385	26.3	270	15.825

The log relatedness odds matrix multipled by 10 is used as a scoring matrix for BSPSM and MDM78; see Table II for BSPSM. The aligned regions listed in this table are ones with the highest score found by 'LFASTA'. Alignment scores of those aligned regions were estimated by using RDF2; each protein pair was shuffled 100 times and alignment scores of both cases were averaged and listed here. The 'ktup' and the cut-off parameters were both set to 1 in the 'LFASTA' and 'RDF2'. Default values are used for other parameters in both cases for MDM78 and BSPSM; the deletion penalty is equal to $-12-4(n-1)$, where n is the number of gaps. Refer to Pearson and Lipman (1988) for the details of these parameters. The codes of sequences are:

OKBO2C:	protein kinase, cAMP-dependent, catalytic chain; bovine (350 a.a.)
TVBE66:	kinase-related transforming protein; varicella zoster virus (393 a.a.)
TVBY8:	cell division control protein 28; yeast (*Saccharomyces cerevisiae*) (298 a.a.)
TVCHMS:	kinase-related transforming protein (mos); chicken (349 a.a.)
TVHUP1:	kinase-related transforming protein (pim-1); human (313 a.a.)
TVHURS:	kinase-related transforming protein (ros-1); human (471 a.a.)
TVMVF6:	kinase-related transforming protein (raf); murine sarcoma virus 3611 (323 a.a.)
TVFFRF:	kinase-related transforming protein (Draf-1); fruit fly (fragment) (291 a.a.)

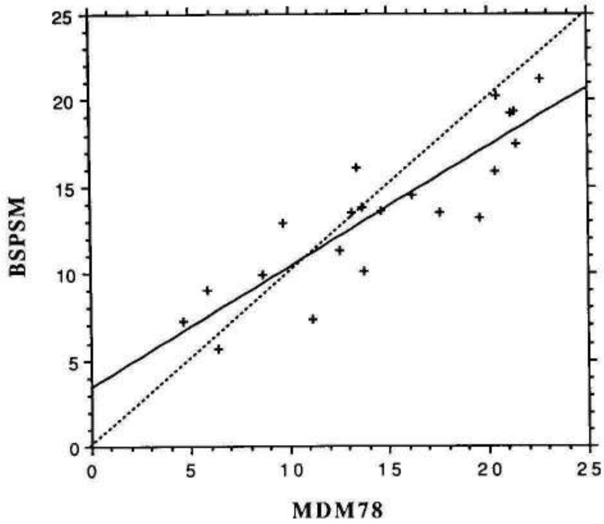

Fig. 6. Comparison of alignment scores in local homology search of the kinase-related transforming protein family. Alignment scores for BSPSM and MDM78, which are listed in Table VI, are plotted against each other. The dotted lines in these figures show the isoscore line. The solid lines are the regression lines of the ordinate on the abscissa. The regression line is $y = 3.41 + 0.69x$, and the correlation coefficient is 0.88.

shown in Table V. Excluding the closely related protein pairs that have trivially large alignment scores, those scores are plotted in Figure 5. Parts (a) and (b) show that BSPSM tends to yield larger alignment scores than UM and GCM, because most of marks are located above the isoscore line shown by the dotted line in those figures. This tendency is remarkable in the region of low alignment score. This indicates that BSPSM can detect more distant relationships than UM and GCM. The situation is not simple in the cases of BSPSM versus SGM and BSPSM versus MDM78. Figure 5(c and d) shows that SGM yields many scores whose values are larger than those of BSPSM. That tendency is more clear in the case of BSPSM versus MDM78 (see Figure 5e and f). However, those figures also show that scores whose values are larger for BSPSM than for SGM or MDM78 tend to be located in the region of low alignment score. The regression lines in Figure 5(c and d) indicate the possibility that BSPSM may yield larger scores than SGM in the region below 4 SD of the SGM score. In the case of BSPSM versus MDM78, that point is between 2 and 4 SD of the MDM78 score, indicating that MDM78 may be superior throughout the whole meaningful region since even 3 SD is a marginal score for saying that two sequences may have a common ancestor. Feng *et al.* (1985) pointed out in their work of testing scoring matrices that many sequences in the hemoglobin superfamily were used to evaluate MDM78, so that MDM78 may be more biased towards the hemoglobin superfamily. They tried to use an alternative data set of a kinase-related transforming protein family. However, this data set does not include many distantly related sequences, making it difficult to compare scores of BSPSM with those of MDM78 in the region of low alignment score. So, we tried carefully to choose distantly related sequences from the kinase-related transforming protein family in a recent PIR protein database.

In the case of the kinase-related transforming protein family, members are not homologous to one another over a whole sequence but have local regions which are homologous. Therefore, a local homology search should be done. For this reason we did not use the 'ALIGN' program to calculate global alignments of a certain region in the proteins, as Feng *et al.* (1985) did, but the 'LFASTA' program to find locally homologous regions between the proteins. Table VI shows alignment scores, the length of aligned regions and the percentage identity in the aligned region for scoring matrices BSPSM and MDM78. The aligned regions shown in Table VI are ones with the highest score found by 'LFASTA'. To calculate alignment scores of the most similar regions found by 'LFASTA', the mean

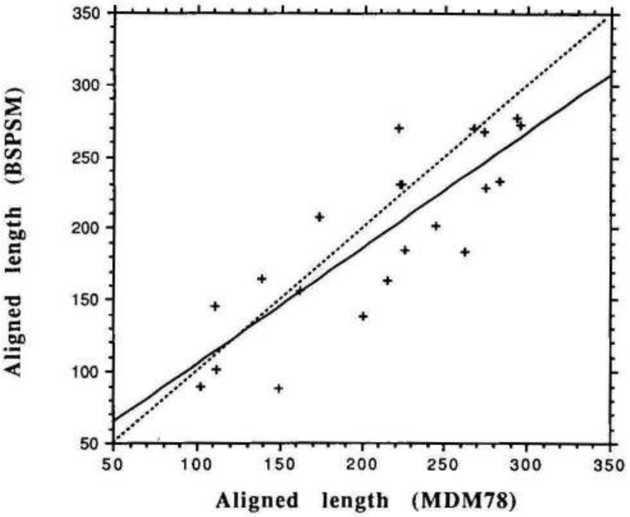

Fig. 7. Comparison of the length of best aligned regions in local homology search of the kinase-related transforming protein family. The lengths of aligned regions for BSPSM and MDM78, which are listed in Table VI, are plotted against each other. The dotted line in this figure shows the iso-length line. The solid line is the regression line of the ordinate on the abscissa. The regression line is $y = 24 + 0.81x$, and the correlation coefficient is 0.83.

and standard deviation of scores in the case of randomized sequences were estimated by using 'RDF2'. To do a complete search, the 'ktup' parameter, which determines how many consecutive identities are required in a match, and the cut-off parameter of scores of initial regions for the optimization step are both set to 1 in the 'LFASTA' and 'RDF2' programs. Default values are used for other parameters for both MDM78 and BSPSM; the deletion penalty is equal to $-12\text{-}4(n\text{-}1)$, where n is the number of gaps. It should be noted here that the log relatedness odds matrix multiplied by 10 is used as a scoring matrix for BSPSM and MDM78 (see Table II for BSPSM). Please refer to Pearson and Lipman (1988) for the details of these parameters. The number of shuffles in 'RDF2' is 100. Each protein is shuffled and alignment scores of both cases are averaged and listed in Table VI. Most aligned regions are <30% identity, indicating that those protein pairs are distantly related. Even so, the alignment scores of those regions are quite high; almost all are >6 SD, indicating that they evolved from a common ancestor. Some regions have alignment scores >15 SD. Please note that values of alignment score by 'RDF2' should not be compared with those of 'ALIGN', because the distribution (Pearson and Lipman, 1988) of scores in randomized sequences may be different in the two methods. Alignment scores for BSPSM are plotted against those for MDM78 in Figure 6, and the length of best aligned regions is shown in Figure 7. In the region of alignment score >15 SD, alignment scores tend to be higher for MDM78 than for BSPSM, but in the region <10 SD BSPSM can yield higher alignment scores than MDM78, indicating that BSPSM may be more useful for detecting distantly related relationships between proteins than MDM78.

Discussion

The average energy increments of protein native structures caused by amino acid exchanges were estimated, and used to evaluate the fitness of amino acid replacements. The estimated values of the average energy increments for amino acid exchanges reasonably represent the physico-chemical similarities of amino acids. Not taken into account in these estimates of the average energy increments is variation in the size of amino acid side chains. Replacement of a small residue in size by a bulky residue could destabilize protein structures. Such an effect should be taken into account; however, the flexibility of protein structures may reduce such volume effects.

Also, similarities in the conformational properties of residues are not explicitly taken into account in the estimate of the average energy increments of amino acid exchanges. As evidence for the unimportance of these effects, Kelly and Holladay (1987) showed that the mean area buried scale of Rose *et al.* (1985) and the optimal matching hydrophobicities (OMHs) scale of Sweet and Eisenberg (1983) are well conserved among homologous sequences as well as other hydrophobicity scales but that the conformational scale of alpha, beta and coil propensities of amino acids compiled by Chou and Fasman (1978) is the least conserved of the scales that they examined. Their result indicates that the hydrophobicity may be more essential for retaining the 3-D structure than the conformational properties of the alpha, beta and coil regions. The contact energies estimated by Miyazawa and Jernigan (1985) include the hydrophobic interaction between amino acids. Rose *et al.* (1985) pointed out that their values of the characteristic fractional area loss and the average energy change of the *i* type of residue, e_{ir}, upon contact formation are well correlated (the correlation coefficient is 0.94). The e_{ir} also correlates well with the optimal matching hydrophobicities of Sweet and Eisenberg (1983) (the correlation coefficient is 0.89).

Sweet and Eisenberg (1983) calculated a set of optimal matching hydrophobicities (OMHs) of the 20 kinds of amino acids which will give the maximum possible value of the correlation coefficient for sequences being compared, from the observed frequency of amino acid replacements compiled by Dayhoff *et al.* (1978). They showed that significant correlations of OMHs are obtained for sequences whose 3-D structures are similar, even though the alignment has few identical residues. Because their scale depends only on a single residue type, some information included in the observed frequency of amino acid replacements may be lost. The average energy change, e_{ir}, upon contact formation cannot reflect the full information of the residue–residue interactions that the contact energies e_{ij} do; it is a kind of average of e_{ij}. The maximum amount of information should be utilized, if available and appropriate.

In this work, the log relatedness odds matrix is used as a scoring matrix to detect distant relationships between protein sequences. The log relatedness odds matrix includes the effects of the genetic code and base mutation rates on amino acid substitutions. The genetic code and base mutation rates must be taken into account for analyzing amino acid substitutions that occurred in the evolutionary process. However, if one is interested in comparing amino acid sequences with each other in order to judge whether the two sequences may adopt similar 3-D structures or not, only the physico-chemical and conformational properties of amino acids should be considered. The amino acid exchange energy matrix ($\Delta\epsilon_{ij}$) rather than the log relatedness odds matrix should be used in such a case to score amino acid matches and mismatches.

Alignment scores strongly depend on the pattern of amino acid substitutions that may vary widely among proteins. This is why GCM and UM yield better results for some proteins than others, although they generally yield significantly poor results for distantly related sequences. This may happen because of the conservation of what are usually more mutable amino acids. In detecting relationships, best results would presumably be obtained with a matrix corresponding to the same evolutionary distance as that between the sequences being compared. UM should therefore yield larger scores for closely related sequences than MDM78 and BSPSM, which are log relatedness odds matrices corresponding to 250 PAM. Of course, we are most interested in obtaining a significant score for comparison between distantly related sequences, and so BSPSM may be useful for such a purpose as well as MDM78, as shown in Figure 6.

One significant advantage of BSPSM over MDM78 is that we can calculate different sets of BSPSM by adjusting the equilibrium frequencies of codons for each protein family, and also by changing base mutation rates. We tried to change the parameter m in equation (14) for transversion and transition. The alignment scores in global homology search are not significantly different, even if m for transition is twice m for transversion. The effects of changing the equilibrium frequencies of codons have not been examined.

References

Chou,P.Y. and Fasman,G.D. (1978) *Annu. Rev. Biochem.*, **47**, 251−276.

Covell,D.G. and Jernigan,R.L. (1990) *Biochemistry*, **29**, 3287−3294.

Dayhoff,M.O. and Eck,R.V. (1968) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure 1968*, Vol. 3. National Biomedical Research Foundation, Silver Spring, MD, pp. 33−41.

Dayhoff,M.O., Eck,R.V. and Park,C.M. (1972) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure 1972*, Vol. 5. National Biomedical Research Foundation Washington, DC, pp. 89−99.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure 1978*, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 345−352.

Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) *J. Mol. Evol.*, **21**, 112−125.

Kelly,L. and Holladay,L.A. (1987) *Protein Engng*, **1**, 137−140.

Kimura,M. (1968) *Nature*, **217**, 624−626.

Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1435−1441.

Matsumura,M., Becketel,W.J. and Matthews,B.W. (1988) *Nature*, **334**, 406−410.

McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409−424.

McLachlan,A.D. (1972) *J. Mol. Biol.*, **64**, 417−437.

Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534−552.

Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443−453.

Orcutt,B.C., Dayhoff,M.O., George,D.C. and Barker,W.C. (1984) *PIR report ALI-1284*. National Biomedical Research Foundation, Washington, DC.

Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444−2448.

Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) *Science*, **229**, 834−838.

Schwartz,R.M. and Dayhoff,M.O. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure 1978*, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 353−358.

Sweet,R.M. and Eisenberg,D. (1983) *J. Mol. Biol.*, **171**, 479−488.

Yutani,K., Ogasahara,K., Tsujita,T. and Sugino,Y. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4441−4444.

Received on September 14, 1992; accepted on November 30, 1992